Title

An Upper Bound for Accuracy of Prediction Using GBLUP.

Journal

PloS one, 11(8)

ISSN

1932-6203

Authors

Karaman, Emre
Cheng, Hao
Firat, Mehmet Z
et al.

Publication Date

2016

DOI

10.1371/journal.pone.0161054

Peer reviewed

PLOS ONE

RESEARCH ARTICLE

An Upper Bound for Accuracy of Prediction Using GBLUP

Emre Karaman[1]*, Hao Cheng[2,3], Mehmet Z. Firat[1], Dorian J. Garrick[2,4], Rohan L. Fernando[2]

1 Department of Animal Science, Faculty of Agriculture, Akdeniz University, 07059 Antalya, Turkey, **2** Department of Animal Science, Iowa State University, 50011 Ames, Iowa, United States of America, **3** Department of Statistics, Iowa State University, 50011 Ames, Iowa, United States of America, **4** Institute of Veterinary, Animal and Biomedical Science, Massey University, Palmerston North, New Zealand

* emrkaraman@gmail.com

Citation: Karaman E, Cheng H, Firat MZ, Garrick DJ, Fernando RL (2016) An Upper Bound for Accuracy of Prediction Using GBLUP. PLoS ONE 11(8): e0161054. doi:10.1371/journal.pone.0161054

Editor: Bamidele O. Tayo, Loyola University Chicago, UNITED STATES

Received: November 30, 2015

Accepted: July 29, 2016

Published: August 16, 2016

Data Availability Statement: We uploaded the data file to Dryad Digital Repository with DOI number (doi:10.5061/dryad.3k8g5).

Funding: RLF acknowledges funding from NIH (http://www.nih.gov/) grant R01GM099992. EK was funded by the Akdeniz University (http://www.akdeniz.edu.tr), Turkey. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

This study aims at characterizing the asymptotic behavior of genomic prediction R^2 as the size of the reference population increases for common or rare QTL alleles through simulations. Haplotypes derived from whole-genome sequence of 85 Caucasian individuals from the 1,000 Genomes Project were used to simulate random mating in a population of 10,000 individuals for at least 100 generations to create the LD structure in humans for a large number of individuals. To reduce computational demands, only SNPs within a 0.1M region of each of the first 5 chromosomes were used in simulations, and therefore, the total genome length simulated was 0.5M. When the genome length is 30M, to get the same genomic prediction R^2 as with a 0.5M genome would require a reference population 60 fold larger. Three scenarios were considered varying in minor allele frequency distributions of markers and QTL, for $h^2 = 0.8$ resembling height in humans. Total number of markers was 4,200 and QTL were 70 for each scenario. In this study, we considered the prediction accuracy in terms of an estimability problem, and thereby provided an upper bound for reliability of prediction, and thus, for prediction R^2. Genomic prediction methods GBLUP, BayesB and BayesC were compared. Our results imply that for human height variable selection methods BayesB and BayesC applied to a 30M genome have no advantage over GBLUP when the size of reference population was small (<6,000 individuals), but are superior as more individuals are included in the reference population. All methods become asymptotically equivalent in terms of prediction R^2, which approaches genomic heritability when the size of the reference population reaches 480,000 individuals.

Introduction

The availability of single nucleotide polymorphism (SNP) marker chips for many species has given rise to the era of genomic prediction (GP). As the name suggests, GP refers to the use of genomic information to predict genetic merit and can be applied to animals [1–4], plants [5–7], or to predisposition to disease in personalized medicine [8]. GP utilizes the phenotypes and

SNP genotypes of a group of individuals (hereafter, called the reference population-RP) to estimate marker effects, which are used to predict breeding values, or yet-to-be observed phenotypes of individuals with genotypes (hereafter called the validation population-VP) [1].

The accuracy of GP is influenced by many factors, such as the method used to estimate marker effects [9, 10], the heritability (h^2) and genetic architecture of the trait [10, 11], and the size (n_R) and structure of the RP [11–16]. Among those, the method, and the size and structure of the RP can be chosen or designed utilizing available knowledge about the heritability and genetic architecture of the trait.

One of the challenges of GP using high-density SNP genotypes is the estimation of SNP effects when the number of individuals comprising RP, n_R, is much smaller than the number of SNPs p, $n_R \ll p$. One approach to address this problem is Bayesian regression, which combines prior information on the vector of SNP effects, $\boldsymbol{\beta}$, with the observed data to estimate all p of the β_js [1]. There are several variations of the Bayesian regression approach, differing in the prior distribution for β_j. A commonly used prior for β_j is a normal distribution with a common variance for all loci: $\beta_j \sim N(0, \sigma^2_\beta)$. This is equivalent to ridge regression, and when the ratio $\sigma^2_\beta/\sigma^2_e$ is known, it can be shown that it becomes best linear unbiased prediction (BLUP) [17]. Due to the relationship of this Bayesian regression to ridge regression and BLUP, it is referred to as Bayesian ridge regression (BRR) or random regression BLUP (RR-BLUP). Genomic predictions obtained from RR-BLUP are identical to those obtained from an animal model (GBLUP), where the numerator relationship matrix is replaced by a genomic relationship matrix (**G**) computed from markers [9, 18–21].

The GBLUP method is popular in GP for three reasons: 1) Since for several decades selection decisions in livestock populations have been routinely made based on BLUP for the animal model [22], GBLUP can easily be used with current computer programs without much effort, 2) for BLUP, theory is available to compute the variance of prediction errors, and 3) many important traits in animals, plants and humans are complex in nature, and are controlled by a large number of small effect genes distributed across the entire genome [23–25], resembling an infinitesimal model [1]. When this assumption does not hold, because a few genes have large effects, or because many genes have no effect, mixture models such as BayesB [1] or BayesC [26, 27] can be used, where the prior for β_j has a point mass at zero with probability π, or has a t or normal distribution with probability $(1 - \pi)$ for BayesB and BayesC, respectively.

The squared correlation between the genetic value (u) and its predicted value ($\hat{u}$) is called the reliability of prediction, and is a measure of prediction accuracy. Goddard [16] and Daetwyler et al. [28] have developed approximations for prediction accuracy utilizing the effective population size (N_e), n_R, h^2, and the effective number of chromosomal segments segregating in the population (M_e). Both approximations were developed assuming complete linkage disequilibrium (LD) between marker-QTL pairs. Goddard et al. [29] extended his earlier approach to address the problem of incomplete LD between markers and QTL. Following that extension, reliability of GP can be approximated with $q^2[n_R h^2/(n_R h^2 + M_e/q^2)]$, where q^2 is the proportion of genetic variance explained by markers. In contrast to simulations, in real applications u of an individual is not observed, and therefore, reliability of prediction cannot be directly computed. Thus, the squared correlation between phenotype (y) and $\hat{u}$ (hereafter, called the R^2) is often used as the measure of prediction accuracy. The approximation for reliability in [29] can be modified (See Appendix 1 in S1 Text) to get an approximation for R^2 as:

$$R^2 \approx h^2_M[n_R h^2_M/(n_R h^2_M + M_e)], \quad (1)$$

where h^2_M is the genomic heritability, the proportion of variance explained by markers [30].

In pedigree-based prediction, heritability is a major determinant of R^2. Using both real data and simulations based on real genotypes, de los Campos et al. [30] investigated the relationship between h^2_M and R^2 from GBLUP for complex traits in humans. They examined different scenarios varying in the distribution of minor allele frequencies (MAFs) of markers and QTL. When both RP and VP included only unrelated individuals, R^2 and h^2_M were 0.071 and 0.737 when markers and QTL had similar MAF distributions, and 0.049 and 0.573 when MAF for QTL were low relative to those for markers. It was concluded that h^2_M is not a good indicator of R^2 when the individuals being predicted are not related to the RP. Instead, the authors proposed $[1-(1-b)^2]h^2$ as an upper bound for R^2, where b is the average regression coefficient of the marker derived relationships on QTL derived relationships. Based on the average value of b estimated for candidates unrelated to the reference population, they concluded that the asymptotic upper bound on the R^2 is about 20% of h^2 for unrelated individuals. This conflicts with approximation Eq (1) where it can be seen that the asymptotic value for R^2 is predicted to be h^2_M.

Taking $N_e = 10{,}000$ [31], the average chromosome length (L) as 1.57 Morgans, and the number of chromosomes (k) as 23, to represent humans, $M_e = \frac{2N_eLk}{log(N_eL)} \approx 7 \times 10^4$ [16], and then from Eq (1) the expected R^2 is 0.037 for $h^2_M = 0.737$ and 0.022 for $h^2_M = 0.573$ when $n_R =$ 5,300 as in [30]. These are close to the R^2 values in [30] for unrelated individuals. In order for R^2 to reach about 90% of h^2_M, approximation Eq (1) suggests that a RP of over half a million individuals is needed. Thus, computer simulation will be used to examine whether the upper bound in [30] holds or if R^2 reaches h^2_M when n_R increases as implied by Eq (1). Further, the suggestion in [30] that variable selection methods may have higher accuracy of prediction than GBLUP for complex traits in humans is examined.

There is growing interest on the optimum structure of the RP [13, 32], accounting for its impact on the relationship between VP and RP individuals [9, 12, 13, 15, 33, 34]. On the other hand, the definition of relatedness in most studies is based on concepts associated with pedigrees, which depend on how deep the pedigree is traced and which measure of relationship is used (e.g. average squared relationship, mean relationship). However, the **G** matrix better reflects genetic similarities between individuals than the numerator relationship matrix computed from pedigree. When the main interest is to rank individuals in the VP according to their predicted genetic values, $\hat{u}$, the use of pairwise relationships between VP and RP can be misleading. In this article, we approach GP as an estimability problem and accordingly provide an upper bound for reliability of prediction, and therefore an upper bound for prediction R^2.

Materials and Methods

Data Sets and Simulation of Genomes

The central objective of this study was to examine the asymptotic behavior of R^2, which requires a large RP. To simulate genotypes that resembles the LD structure in humans for a large number of individuals, haplotypes of 85 Caucasian individuals from the 1,000 Genome Project [35] (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/) were used to generate a population of size 10,000. This was accomplished by randomly sampling 20,000 gametes from phased paternal and maternal sequence accounting for crossing over and mutation. Following this, 10,000 offspring were sampled from random mating of 10,000 parents for 111 non-overlapping generations. Generations 101 and 111 were used to form RP and VP, respectively. A mutation rate of 1×10^{-8} was used. Mutations switched the original allele state from 0 to 1, or vice versa. Simulation of the genomes were performed using XSim software, which uses an algorithm that tracks only the positions of crossing over and mutation as well as the origin

Table 1. Percentage of SNPs in a particular range of MAF.

Scenario	MAF	Type	<3%	3%-5%	5%-10%	10%-15%	>15%
SHL	High	Marker	0.061	0.046	0.117	0.114	0.663
	Low	QTL	0.314	0.243	0.229	0.200	0.014
SHR	High	Marker	0.065	0.049	0.119	0.115	0.652
	Random	QTL	0.029	0.057	0.100	0.100	0.714
SRR	Random	Marker	0.082	0.050	0.096	0.104	0.668
	Random	QTL	0.100	0.029	0.114	0.114	0.643

MAF: minor allele frequency; SHL: markers and QTL were selected for high and low MAFs, respectively; SHR: markers were selected for high MAF, whereas QTL were selected at random; SRR: markers and QTL were selected at random

doi:10.1371/journal.pone.0161054.t001

of each chromosomal segment throughout the generations, and finally, drops sequence variants from the founders to the individuals in the last generation [36].

To reduce computational demands, only SNPs with known positions within a 0.1M region of each of the first 5 chromosomes were used. Markers with MAF <0.005 were discarded resulting in a data set including 36,242 SNPs. To make our data set comparable with [30], we scaled down these loci to 4,200 to be used as markers and an additional 70 to represent QTL. Hence, the density of markers (84/cM) and ratio of the number of QTL (n_{QTL}) to markers (1/6) are similar to those in [30] that was based on 300,000 markers and 5,000 QTL. Three different scenarios (SHL-SRR) were created varying in the MAF distributions: (SHL) markers and QTL were selected for high and low MAFs, respectively; (SHR) markers were selected for high MAF, whereas QTL were selected at random; and (SRR) markers and QTL were selected at random. Hence, among the 36,242 SNPs, 4,200 were taken to be markers, and 70 as QTL among those with high, random or low MAF as appropriate. Scenario SHL corresponds to Low-MAF, whereas SHR corresponds to RAND in [30]. The MAF distributions for the three scenarios are summarized in Table 1. Different MAF scenarios allowed us to examine the impact of the structure of SNP chips on accuracy.

Reference and Validation Populations

In this study, RP and VP were separated for 10 generations to ensure that close relatives of the VP individuals do not exist in RP. Both generations consisted of 10,000 individuals. Among the individuals in generation 101, n_R individuals were randomly selected to form a RP, for a range of n_R (75, 150, 500, 1,000, 2,000, 4,000 and 8,000), while n_V = 2,000 individuals were selected among the individuals in generation 111 to form a VP.

Simulation of Phenotypes

Similar to de los Campos et al. [30], a quantitative trait corresponding to human height with heritability 0.8 was simulated. The effects of QTL, α_js, were sampled from a normal distribution, $\alpha_j \sim N(0, 1)$. Since the QTL effect sizes vary by replicate, the genetic variance can vary by replicate. To keep the heritability constant across replicates, therefore, QTL effects were scaled at each replicate. The product of the scaled QTL effects and the QTL genotypes was used to obtain the genetic value of individual i as follows:

$$u_i = \sum_{j=1}^{n_{QTL}} \alpha_j \times Q_{ij},$$

where n_{QTL} is the number of QTL, α_j is the additive effect of j'th QTL, and Q_{ij} is the genotype of individual i at the j'th QTL. In each replicate of each scenario, the same SNPs were designated as markers or QTL, however, QTL effects were separately randomly simulated in each replicate. A standard normal deviate (e_i) was added to each individual's u_i to form its phenotypic value (y_i) with the desired heritability.

Estimation of Marker Effects

The statistical model fitted to the data is:

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + e_i,$$

where y_i is the phenotypic value of individual i in the RP, μ is the overall mean, p is the number of marker loci, x_{ij} is the marker genotype of individual i at locus j, β_j is the allele substitution effect of marker j, and e_i is the random environmental effect assumed to be normally distributed, $e_i \sim N(0,\ \sigma_e^2)$.

To predict the genetic value of individuals, marker effects were first estimated from RP data using BayesB and BayesC methods, which differ in the prior assumptions for marker effects as described previously. BayesC with $\pi = 0$ is identical to GBLUP when σ_β^2 is treated as unknown with a scaled inverse chi-square prior. In the BayesB and BayesC analyses, π was 0.98, whereas GBLUP results were obtained using BayesC with $\pi = 0$. A total of 11,000 Markov chain Monte Carlo iterations were used, with the first 1,000 excluded as the burn-in. Marker effects were estimated from separate analyses with inclusion or exclusion of QTL from the marker panel. Analyses were performed using GenSel software [37, 38]. In order to evaluate how frequently a marker was included in the model in a BayesB or BayesC run, the model frequency (MF) in GenSel output can be used which is defined as the proportion of iterations or models that included that marker.

Prediction of Genetic Values, and Prediction Accuracy

Given the estimates of the marker effects, the u of the VP individuals was predicted as:

$$\hat{u}_i = \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$$

where $\hat{\beta}_j$ is the estimated effect of locus j, and x_{ij} is the marker genotype of i'th individual at locus j. The prediction R^2 was calculated as the squared correlation between the phenotypes, y_i, and $\hat{u}_i$ of VP individuals.

Using the regression model:

$$\mathbf{g}_{M,i} = b_i \times \mathbf{g}_{Q,i} + \epsilon_i \quad (2)$$

where $\mathbf{g}_{M,i}$ is the vector of marker derived relationships of i'th individual in VP to all the individuals in RP, and $\mathbf{g}_{Q,i}$ is the vector of QTL derived relationships, and ϵ_i is a vector of residuals, the regression coefficient, b_i, was estimated [30] and averaged across all individuals in VP for each replicate. Both forms of relationships were obtained from the realized relationship matrix, $\mathbf{G}$:

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{p}$$

where $\mathbf{Z}$ is the matrix of genotypes constructed by the standardized vector of (marker or QTL) genotypes including RP and VP, and p is the number of loci (marker or QTL) [25]. Standardization of genotypes was done as follows, where $\mathbf{x}_j$ is a vector of genotypes of individuals at j'th loci, and q_j is the allele frequency:

$$\mathbf{z}_j = \frac{\mathbf{x}_j - 2q_j}{\sqrt{2q_j(1-q_j)}}.$$

Scenarios SHL, SHR, and SRR, were replicated 10 times, and results were averaged across replicates.

A Connection between Prediction Accuracy and Estimability in a Fixed Linear Model

Consider the linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\boldsymbol{\beta}$ is assumed fixed. In this setting, a linear function, $\mathbf{k}'\boldsymbol{\beta}$ is said to estimable only if the estimator $\mathbf{k}'\hat{\boldsymbol{\beta}}$ has expected value $\mathbf{k}'\boldsymbol{\beta}$. When $\hat{\boldsymbol{\beta}}$ is the least-squares estimator, it is known that $\mathbf{k}'\boldsymbol{\beta}$ is estimable only when $\mathbf{k}'$ is a linear function of the rows of $\mathbf{X}$ [39].

In RR-BLUP, $\boldsymbol{\beta}$ is considered random with null mean, and the BLUP of $\mathbf{x}'_V\boldsymbol{\beta}$ representing an individual in VP with genotypes $\mathbf{x}_V$ is unbiased in the sense that

$$E(\mathbf{x}'_V\tilde{\boldsymbol{\beta}}) = E(\mathbf{x}'_V\boldsymbol{\beta}) = 0,$$

where $\tilde{\boldsymbol{\beta}}$ is the BLUP of $\boldsymbol{\beta}$. To see the connection between prediction accuracy and estimability, let $\mathbf{X}_R$ be the genotype matrix of RP and let $\Re(\mathbf{X}_R)$ denote the row space of $\mathbf{X}_R$. Then, any vector $\mathbf{x}_V$ can be written as the sum of two vectors: $\mathbf{x}_{V_1} = \mathbf{Q}_{X'_R}\mathbf{x}_V$, which is in $\Re(\mathbf{X}_R)$, and $\mathbf{x}_{V_2} = (\mathbf{I} - \mathbf{Q}_{X'_R})\mathbf{x}_V$, which is orthogonal to $\Re(\mathbf{X}_R)$, i.e., the vector of validation genotypes can be written as

$$\mathbf{x}_V = \mathbf{x}_{V_1} + \mathbf{x}_{V_2}, \tag{3}$$

where $\mathbf{Q}_{X'_R} = \mathbf{X}'_R(\mathbf{X}_R\mathbf{X}'_R)^{-}\mathbf{X}_R$ [39]. From Eq (3), $\mathbf{x}'_V\tilde{\boldsymbol{\beta}}$ of an individual in VP can be partitioned as:

$$\mathbf{x}'_V\tilde{\boldsymbol{\beta}} = \mathbf{x}'_{V_1}\tilde{\boldsymbol{\beta}} + \mathbf{x}'_{V_2}\tilde{\boldsymbol{\beta}}. \tag{4}$$

It is shown in Appendix 2 of S1 Text that $\tilde{\boldsymbol{\beta}}$ is in $\Re(\mathbf{X}_R)$, and therefore, the BLUP of $\mathbf{x}'_V\boldsymbol{\beta}$ is

$$\mathbf{x}'_V\tilde{\boldsymbol{\beta}} \quad = \quad \mathbf{x}'_{V_1}\tilde{\boldsymbol{\beta}}. \tag{5}$$

Accordingly, $\mathbf{x}'_{V_2}\tilde{\boldsymbol{\beta}} = 0$, which is the mean of its prior and does not depend on the data. This can be seen more clearly by writing the correlation between $u_i = \mathbf{x}'_V\boldsymbol{\beta}$ and $\hat{u}_i = \mathbf{x}'_V\tilde{\boldsymbol{\beta}}$ in terms of Eq (5) as shown below. Under BLUP assumptions [22],

$$Cor(u_i, \hat{u}_i) \quad = \quad \sqrt{\frac{Var(\hat{u}_i)}{Var(u_i)}}, \tag{6}$$

and from Eq (5), the numerator of Eq (6) becomes $Var(\mathbf{x}'_{V_1}\tilde{\boldsymbol{\beta}})$. So, Eq (6) can be written as

$$Cor(u_i, \hat{u}_i) = \sqrt{\frac{Var(\mathbf{x}'_{V_1}\tilde{\boldsymbol{\beta}})}{Var(\mathbf{x}'_{V}\boldsymbol{\beta})}}.$$

Clearly, $Var(\mathbf{x}'_{V_2}\tilde{\boldsymbol{\beta}}) = 0$ and does not contribute to $Cor(u_i, \hat{u}_i)$. An individual for whom the genotype vector is orthogonal to all genotypes in RP ($\mathbf{x}_V = \mathbf{x}_{V_2}$) can be thought of as being genomically unrelated to the RP. For such an individual, $Cor(u_i, \hat{u}_i)$ would be zero. On the other hand, an individual for whom the genotype vector is in the row space of genotypes in RP ($\mathbf{x}_V = \mathbf{x}_{V_1}$) can be thought of as having a perfect genomic relationship to the RP. This is not a genomic relationship between two individuals, and it does not require a perfect or even high relationship with any individual in RP. For such an individual, $Cor(u_i, \hat{u}_i)$ will approach 1 as $Var(\mathbf{x}'_{V_1}\tilde{\boldsymbol{\beta}})$ approaches $Var(\mathbf{x}'_{V}\boldsymbol{\beta})$. Generally, the maximum value of $Cor(u_i, \hat{u}_i)$ is $\sqrt{\frac{Var(\mathbf{x}'_{V_1}\boldsymbol{\beta})}{Var(\mathbf{x}'_{V}\boldsymbol{\beta})}}$, which is the square root of reliability defined as $Cor^2(u_i, \hat{u}_i)$. Thus

$$UP_i = \frac{\mathbf{x}'_{V_1}\mathbf{x}_{V_1}}{\mathbf{x}'_{V}\mathbf{x}_{V}}$$

is a measure of the upper bound for reliability (See Appendix 3 in S1 Text). When $UP_i = 0$, the reliability of prediction will be zero regardless of the size of RP, and when $UP_i = 1$, the reliability will approach 1 as the size of RP increases. On the other hand when $UP_i < a$, reliability will be less than a regardless of the size of the RP. In addition, $h^2\ UP_i$ is the upper bound of the R^2 for individual i.

In order to examine the utility of this upper bound, we carried out a simulation study randomly selecting 5,000 of the 36,242 available SNPs. All of the selected SNPs were used as QTL, but two quantitative traits were simulated representing $h^2 = 0.8$ or 0.999. Among the 10,000 individuals in generation 101, n_R individuals were randomly selected to form a RP, where n_R was varied (500, 1,000, 2,000 and 5,000), while $n_V = 2{,}000$ individuals were selected among the individuals of generation 111 to form a VP. The GBLUP method was used to predict genetic values, again through BayesC with $\pi = 0$. Other steps of the analysis were the same as given before. Setting $h^2 = 0.999$, allowed us to minimize the estimation error of marker effects, so that prediction accuracy was determined almost entirely by estimability of the genotypes of VP. Effect of relationship to the RP for VP individuals was investigated utilizing the maximum relationships only at QTL level, $max(\mathbf{g}_{V_i})$, of individual i in VP to the individuals in RP. Individuals in VP were classified into a low (L) and high (H) relationship group: when an individual's maximum relationship was lower than 0.15, it was assigned to L, while an individual with maximum relationship equal or greater than 0.25 was assigned to H.

Results and Discussion

R^2 for GBLUP

Prediction R^2s obtained using the GBLUP method with markers only are summarized in Fig 1 for varying RP sizes and all the scenarios along with the estimated heritabilities and the upper bound suggested by [30]. It is clear from the figure that an increase in the number of individuals in RP results in an increase in prediction R^2 in all the scenarios. This is true even when the markers and QTL have an opposite MAF distribution (SHL), which may be the case in real data studies. In SHL, when RP involves only 75 individuals R^2 was 0.017, while a maximum R^2 of 0.569 was obtained for the largest RP size of 8,000 (Table 2). In SHR, where the markers had

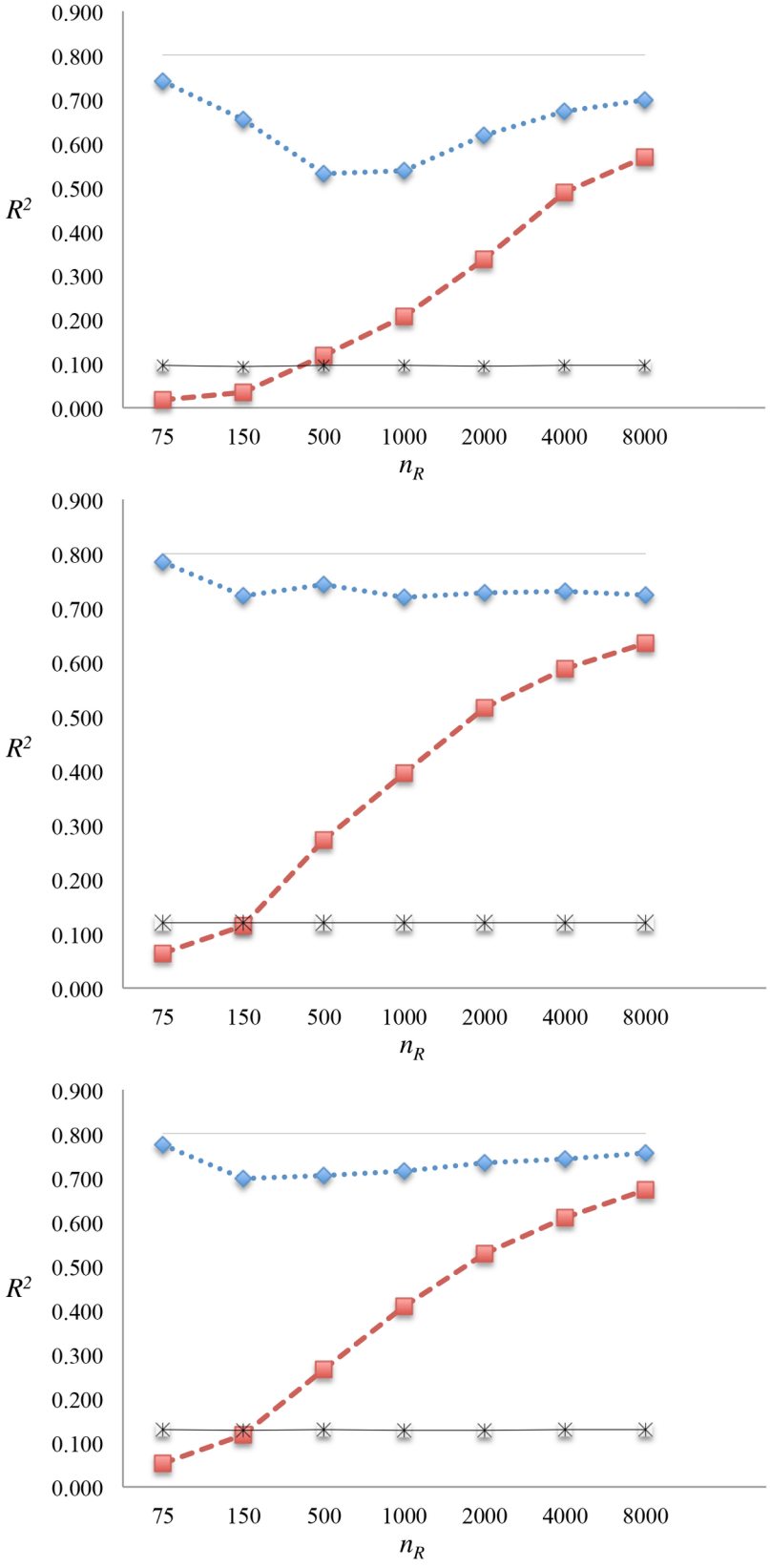

0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
R^2
75
150
500
1000
2000
4000
8000
n_R
(SHL)
(SHR)
(SRR)

Fig 1. Summary of GBLUP results. SHL: markers and QTL were selected for high and low MAFs, respectively; SHR: markers were selected for high MAF, whereas QTL were selected at random; SRR: markers and QTL were selected at random; n_R: number of individuals in RP; R^2: squared correlation between y and $\hat{u}$; prediction R^2 of GBLUP (red line); genomic heritability, h^2_M (blue line); simulated heritability, h^2_{sim} (grey line); the upper bound suggested in [30], $[1 - (1 - b)^2]h^2$ (black line).

doi:10.1371/journal.pone.0161054.g001

Table 2. Results of SHL.

Method	n_R	n_V	R^2	h^2_M	R^2_{M+QTL}	h^2_{M+QTL}
GBLUP	75	2,000	0.017(0.004)	0.741(0.021)	0.023(0.005)	0.686(0.027)
BayesB	75	2,000	0.032(0.007)	0.731(0.013)	0.069(0.016)	0.770(0.015)
BayesC	75	2,000	0.027(0.005)	0.708(0.027)	0.052(0.012)	0.722(0.030)
GBLUP	150	2,000	0.035(0.004)	0.654(0.030)	0.046(0.005)	0.661(0.029)
BayesB	150	2,000	0.082(0.015)	0.664(0.021)	0.174(0.022)	0.730(0.019)
BayesC	150	2,000	0.062(0.011)	0.650(0.032)	0.130(0.016)	0.721(0.027)
GBLUP	500	2,000	0.119(0.011)	0.531(0.037)	0.159(0.011)	0.618(0.035)
BayesB	500	2,000	0.371(0.021)	0.605(0.018)	0.660(0.009)	0.788(0.007)
BayesC	500	2,000	0.382(0.026)	0.635(0.019)	0.683(0.008)	0.801(0.007)
GBLUP	1,000	2,000	0.207(0.010)	0.539(0.021)	0.283(0.009)	0.640(0.017)
BayesB	1,000	2,000	0.486(0.018)	0.604(0.017)	0.749(0.004)	0.794(0.003)
BayesC	1,000	2,000	0.492(0.019)	0.610(0.015)	0.755(0.003)	0.795(0.003)
GBLUP	2,000	2,000	0.336(0.012)	0.617(0.010)	0.465(0.008)	0.734(0.008)
BayesB	2,000	2,000	0.571(0.011)	0.628(0.008)	0.784(0.002)	0.801(0.002)
BayesC	2,000	2,000	0.573(0.011)	0.629(0.007)	0.786(0.002)	0.801(0.002)
GBLUP	4,000	2,000	0.489(0.008)	0.673(0.006)	0.619(0.002)	0.773(0.002)
BayesB	4,000	2,000	0.602(0.009)	0.652(0.007)	0.789(0.003)	0.802(0.001)
BayesC	4,000	2,000	0.599(0.009)	0.649(0.007)	0.789(0.003)	0.801(0.001)
GBLUP	8,000	2,000	0.569(0.008)	0.698(0.005)	0.702(0.002)	0.792(0.001)
BayesB	8,000	2,000	0.624(0.008)	0.659(0.006)	0.794(0.002)	0.801(0.001)
BayesC	8,000	2,000	0.622(0.008)	0.656(0.006)	0.794(0.002)	0.801(0.001)

SHL: markers and QTL were selected for high and low MAFs, respectively; n_R: number of individuals in RP; n_V: number of individuals in VP; R^2 and h^2_M are the predictive ability and genomic heritability when QTL are not in the panel; R^2_{M+QTL} and h^2_{M+QTL} are the predictive ability and genomic heritability when QTL are in the panel

doi:10.1371/journal.pone.0161054.t002

high MAF distribution, and QTL were selected completely at random, a RP size of 75 resulted an R^2 of 0.064, while a RP of size 8,000 yielded the highest R^2 of 0.635 (Table 3). When the markers and QTL were selected completely at random, R^2s were higher than their counterparts in SHL (Table 4). The lowest and highest R^2 values in SRR were 0.053 and 0.673, for a RP size of 75 and 8,000, respectively.

Heritability estimates were volatile in SHL, but almost flat in SHR and SRR. Moreover, the estimates of heritability were always greater for SHR and SRR than for SHL. In all scenarios, the R^2 of GBLUP increased towards h^2_M. It is very likely that when sequence data are used, the fitted genotypes include the QTL. This motivated analyses with inclusion of the QTL genotypes in the marker panel, and even in that case, the predictive accuracy of GBLUP could not attain the estimated heritability for RP sizes considered here. However, the trend suggests that the predictive accuracy of GBLUP could achieve the heritability at a sufficiently large RP size. Regression coefficients for the marker derived relationships on the QTL derived relationships

Table 3. Results of SHR.

Method	n_R	n_V	R^2	h^2_M	R^2_{M+QTL}	h^2_{M+QTL}
GBLUP	75	2,000	0.064(0.008)	0.784(0.011)	0.087(0.009)	0.757(0.013)
BayesB	75	2,000	0.051(0.005)	0.763(0.014)	0.092(0.014)	0.790(0.013)
BayesC	75	2,000	0.065(0.008)	0.754(0.015)	0.106(0.015)	0.775(0.014)
GBLUP	150	2,000	0.115(0.012)	0.722(0.028)	0.155(0.014)	0.750(0.023)
BayesB	150	2,000	0.207(0.020)	0.708(0.019)	0.298(0.021)	0.775(0.018)
BayesC	150	2,000	0.209(0.020)	0.729(0.027)	0.296(0.023)	0.793(0.022)
GBLUP	500	2,000	0.272(0.011)	0.743(0.013)	0.353(0.010)	0.819(0.009)
BayesB	500	2,000	0.473(0.014)	0.681(0.005)	0.682(0.009)	0.805(0.007)
BayesC	500	2,000	0.473(0.014)	0.692(0.005)	0.683(0.009)	0.808(0.007)
GBLUP	1,000	2,000	0.395(0.007)	0.719(0.012)	0.494(0.006)	0.803(0.008)
BayesB	1,000	2,000	0.569(0.009)	0.673(0.010)	0.754(0.003)	0.804(0.003)
BayesC	1,000	2,000	0.571(0.009)	0.675(0.010)	0.753(0.003)	0.803(0.003)
GBLUP	2,000	2,000	0.516(0.013)	0.728(0.008)	0.616(0.008)	0.802(0.004)
BayesB	2,000	2,000	0.629(0.011)	0.691(0.008)	0.780(0.003)	0.804(0.002)
BayesC	2,000	2,000	0.629(0.011)	0.691(0.008)	0.780(0.003)	0.803(0.002)
GBLUP	4,000	2,000	0.588(0.007)	0.730(0.005)	0.686(0.003)	0.800(0.002)
BayesB	4,000	2,000	0.655(0.009)	0.697(0.006)	0.789(0.002)	0.802(0.001)
BayesC	4,000	2,000	0.654(0.008)	0.696(0.006)	0.789(0.002)	0.802(0.001)
GBLUP	8,000	2,000	0.635(0.016)	0.724(0.011)	0.739(0.003)	0.800(0.001)
BayesB	8,000	2,000	0.665(0.018)	0.691(0.013)	0.796(0.002)	0.800(0.001)
BayesC	8,000	2,000	0.665(0.018)	0.690(0.013)	0.796(0.002)	0.800(0.001)

SHR: markers were selected for high MAF, whereas QTL were selected at random; n_R: number of individuals in RP; n_V: number of individuals in VP; R^2 and h^2_M are the predictive ability and genomic heritability when QTL are not in the panel; R^2_{M+QTL} and h^2_{M+QTL} are the predictive ability and genomic heritability when QTL are in the panel

doi:10.1371/journal.pone.0161054.t003

were obtained for every individual in VP, so that the upper bound for prediction R^2, $[1-(1-b)^2]h^2$, suggested by [30] could be plotted for varying RP sizes as in Fig 1. The average of the regression coefficients (results not given), b, were almost invariant to RP size (varying only after 2nd digit), thereby yielding an upper bound that was invariant despite the predictive accuracy increasing with RP size. These results together demonstrate $[1-(1-b)^2]h^2$ is not an upper bound for R^2 as claimed in [30]. When a sufficient number of individuals is in RP, predictive accuracy of GBLUP can reach h^2_M.

Using the Formula (1), the asymptotic value of R^2 reaches the genomic heritability. In this simulation study, the predicted values of R^2 for SHL-SRR from Eq (1) are 0.554, 0.579 and 0.611 with n_R = 8,000, $M_e = \frac{2N_eLk}{log(N_eL)} = 1,448$, for h^2_M of 0.698, 0.724, 0.757, respectively. These predicted values are lower than the predictive accuracies using only markers (0.569, 0.635 and 0.673 for SHL-SRR, respectively), or the predictive accuracies when QTL were also in the panel (0.702, 0.739 and 0.740 for SHL-SRR, respectively).

Previous studies have shown that predictive accuracy increases with an increase in the RP size [11, 15, 16]. Our results are in line with previous findings with ~10 to 35 fold increase in R^2 using only markers when the RP size was increased from 75 to 8,000. Meuwissen [40], suggested the use of large RPs in estimation of marker effects, particularly for the GBLUP. However, in [30], the predictive accuracy of GLUP method was assessed using a small number of individuals (n_R of 5,300) relative to the 300,000 markers fitted in the model. Even though there are many other factors influencing predictive accuracy, a possible explanation of the low R^2

Table 4. Results of SRR.

Method	n_R	n_V	R^2	h^2_M	R^2_{M+QTL}	h^2_{M+QTL}
GBLUP	75	2,000	0.053(0.008)	0.775(0.009)	0.072(0.010)	0.748(0.009)
BayesB	75	2,000	0.073(0.016)	0.768(0.009)	0.124(0.021)	0.789(0.008)
BayesC	75	2,000	0.077(0.015)	0.753(0.010)	0.123(0.021)	0.774(0.009)
GBLUP	150	2,000	0.117(0.012)	0.698(0.041)	0.157(0.013)	0.724(0.037)
BayesB	150	2,000	0.187(0.027)	0.724(0.028)	0.304(0.024)	0.799(0.018)
BayesC	150	2,000	0.190(0.026)	0.731(0.044)	0.307(0.024)	0.814(0.023)
GBLUP	500	2,000	0.266(0.011)	0.705(0.014)	0.345(0.011)	0.790(0.009)
BayesB	500	2,000	0.501(0.017)	0.672(0.010)	0.691(0.006)	0.796(0.006)
BayesC	500	2,000	0.499(0.016)	0.682(0.010)	0.691(0.007)	0.800(0.006)
GBLUP	1,000	2,000	0.409(0.009)	0.716(0.017)	0.505(0.008)	0.801(0.011)
BayesB	1,000	2,000	0.588(0.008)	0.679(0.013)	0.760(0.004)	0.803(0.007)
BayesC	1,000	2,000	0.587(0.008)	0.682(0.013)	0.761(0.004)	0.803(0.007)
GBLUP	2,000	2,000	0.528(0.011)	0.735(0.009)	0.624(0.006)	0.803(0.004)
BayesB	2,000	2,000	0.647(0.013)	0.699(0.010)	0.784(0.002)	0.803(0.002)
BayesC	2,000	2,000	0.647(0.013)	0.698(0.010)	0.784(0.002)	0.802(0.002)
GBLUP	4,000	2,000	0.609(0.007)	0.743(0.005)	0.693(0.004)	0.800(0.001)
BayesB	4,000	2,000	0.678(0.007)	0.709(0.006)	0.792(0.002)	0.801(0.001)
BayesC	4,000	2,000	0.678(0.007)	0.708(0.005)	0.792(0.002)	0.801(0.001)
GBLUP	8,000	2,000	0.673(0.005)	0.757(0.003)	0.740(0.003)	0.801(0.001)
BayesB	8,000	2,000	0.706(0.006)	0.729(0.004)	0.795(0.002)	0.801(0.001)
BayesC	8,000	2,000	0.705(0.005)	0.728(0.004)	0.795(0.002)	0.801(0.001)

SRR: markers and QTL were selected at random; n_R: number of individuals in RP; n_V: number of individuals in VP; R^2 and h^2_M are the predictive ability and genomic heritability when QTL are not in the panel; R^2_{M+QTL} and h^2_{M+QTL} are the predictive ability and genomic heritability when QTL are in the panel

doi:10.1371/journal.pone.0161054.t004

given in [30] might be the small RP size. When n_R was 75 and the number of markers to be estimated was 4,200, predictive accuracies in scenarios SHL and SHR were low (0.017 in SHL, and 0.064 in SHR) as in the corresponding MAF scenarios in [30] for GENOVA data set. However, as mentioned above, predictive accuracy increased with the inclusion of more individuals in RP.

In derivation of the upper bound for R^2 in [30], the conditional expectation of genetic values of VP individuals was written as

$$E(u_{V_i} \mid \mathbf{y}_R) = \mathbf{g}_{Q,i}[\mathbf{G}_Q\sigma^2_u + \mathbf{I}\sigma^2_e]^{-1}\mathbf{y}_R, \quad (7)$$

where $\mathbf{G}_Q$ is the relationship matrix of RP at the QTL level, and $\mathbf{y}_R$ is the vector of centered phenotypes of RP. The QTL genotypes of individuals are not known in reality, therefore genomic relationships are computed from marker genotypes instead of QTL. Thus, the conditional expectation is approximated as

$$E(u_{V_i} \mid \mathbf{y}_R) \approx \mathbf{g}_{M,i}[\mathbf{G}_M\sigma^2_u + \mathbf{I}\sigma^2_e]^{-1}\mathbf{y}_R, \quad (8)$$

where $\mathbf{G}_M$ is the relationship matrix of RP at the marker level. In the derivation of an upper bound for R^2 when this approximation is used, however, it was assumed that $\mathbf{G}_Q$ was known, and $\mathbf{g}_{M,i}$ was written as $\mathbf{g}_{M,i} = b_i\,\mathbf{g}_{Q,i} + \epsilon_i$ [30]. Then, as explained in [30], the approximation can be expressed as

$$E(u_{V_i} \mid \mathbf{y}_R) \approx b_i\mathbf{g}_{Q,i}[\mathbf{G}_Q\sigma^2_u + \mathbf{I}\sigma^2_e]^{-1}\mathbf{y}_R. \quad (9)$$

In a population of unrelated individuals, the expected value of genomic relationships will be zero. When genomic relationships, $\mathbf{g}_{M,i}$ and $\mathbf{g}_{Q,i}$, are computed using 300,000 markers and 5,000 QTL as in [30], $\mathbf{g}_{Q,i}$ will have a much larger variance than $\mathbf{g}_{M,i}$. This results in the slope, b, of the regression of $\mathbf{g}_{M,i}$ on $\mathbf{g}_{Q,i}$ to be small. Thus, Eq (9) will have a much lower R^2 than Eq (7). This would not be the case if Eq (8) was used, where both $\mathbf{g}_{Q,i}$ and $\mathbf{G}_Q$ are replaced with their marker based counterparts. This can be demonstrated by writing $\mathbf{g}_{M,i} = b\,\mathbf{g}_{Q,i}$ and $\mathbf{G}_M = b\,\mathbf{G}_Q$, where b is the average value of b_i. Then the approximation Eq (8) becomes

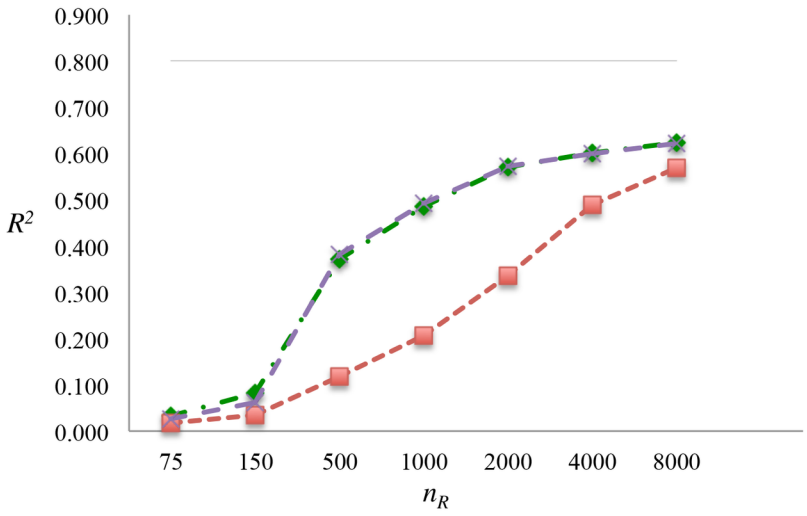
$$E(u_{V_i} \mid \mathbf{y}_R) \approx \mathbf{g}_{Q,i}\left[\mathbf{G}_Q\sigma_u^2 + \mathbf{I}\frac{\sigma_e^2}{b}\right]^{-1}\mathbf{y}_R. \tag{10}$$

This approximation Eq (10) for the conditional expectation is almost identical to Eq (7), and therefore, the R^2 from Eq (10) will be similar to the R^2 from Eq (7).

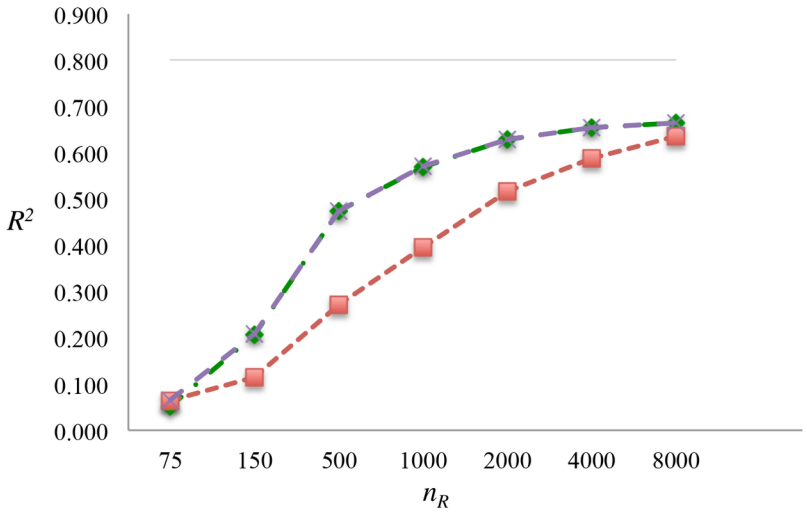
Comparison of Different Methods of Genomic Prediction

Fig 2 shows R^2s of GBLUP, BayesB and BayesC methods for varying values of n_R in all the scenarios when only the markers are in the panel. When $n_R = 75$, all methods had a similar R^2, i.e., the variable selection methods, BayesB and BayesC, had no advantage over GBLUP. As n_R increased, initially, variable selection methods became superior to GBLUP, but eventually all methods yielded similar R^2 values when $n_R = 8{,}000$.

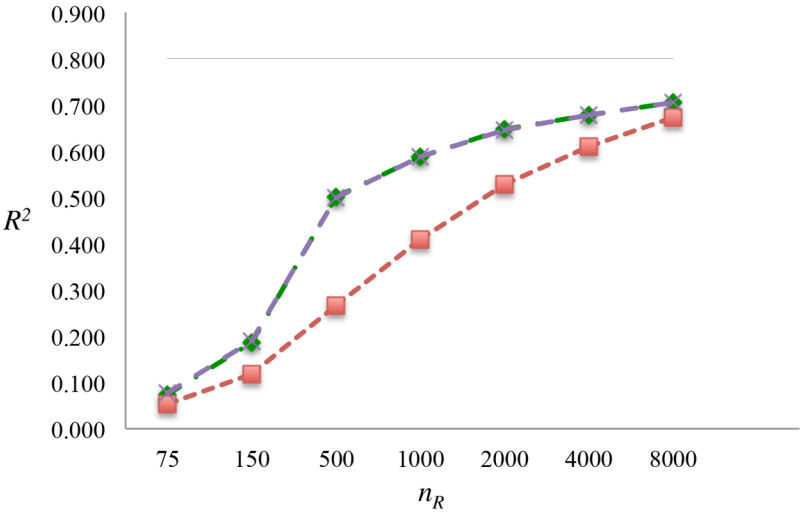
For a given n_R and h^2, predictive accuracy of GBLUP were shown to be highly dependent on M_e, whereas predictive accuracy of BayesB is also dependent on n_{QTL} [11]. An approximation for the reliability of GP with BayesB was suggested with the modification of the equation in [28], which is given as $n_R\,h^2/[n_R\,h^2 + min(n_{QTL}, M_e)]$ [11]. When $n_{QTL} < M_e$, the advantage of variable selection methods, BayesB and BayesC, is expected to be more apparent since they select a subset of loci with an effect on the trait of interest instead of estimating the M_e parameters regardless of whether they have an effect [11]. Using the formula in [11], for $n_R = 75$, $h^2 = 0.8$, $n_{QTL} = 70$ and $M_e = 1{,}448$, the predicted values of R^2 were 0.370 and 0.032 for BayesB and GBLUP, respectively. However, the observed predictive accuracies for BayesB and BayesC in scenarios SHL to SRR were much lower (0.027-0.077) than these predicted values when $n_R = 75$. Fig 3 depicts the MFs of markers in one replicate of SHL for BayesB method when $n_R = 75$ and $n_R = 8{,}000$. It is clear that when $n_R = 75$ MFs followed a uniform distribution with none of the markers having MF higher than 0.4. On the other hand, there are many such markers $0.4 <$ MF when $n_R = 8{,}000$, and as n_R increased from 75 to 8,000, variance of the MF of markers increased more than 100-fold from 1.21×10^{-4} to 214×10^{-4} indicating that when n_R is small, variable selection does not effectively discriminate between markers that are in LD with QTL from those that are not. Therefore, one can not take full advantage of variable selection methods if n_R is not sufficiently large. An increase in n_R from 75 to 8,000 resulted in about a 10- to 20-fold increase in R^2 for BayesB and BayesC (Tables 2–4). On the other hand, the advantage of variable selection methods over GBLUP diminished when $n_R = 8{,}000$, and all methods yielded similar predictive accuracies for this high heritability trait. Variable selection methods shrink marker effects that are very small towards zero, and therefore, these loci do not contribute to the estimation of u. However, when π of BayesC is set to zero, effects of all markers are estimated regardless of their size, which usually add only noise to the estimation of u. On the other hand, when sufficiently large RP is used, the effect of those markers that are very small can be estimated accurately, yielding a high predictive accuracy. This can explain why predictive accuracy from variable selection methods are higher than GBLUP when number of markers are much larger than n_R, and why eventually all methods yielded same predictive accuracies.


0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
R^2
75
150
500
1000
2000
4000
8000
n_R

(SHL)


0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
R^2
75
150
500
1000
2000
4000
8000
n_R

(SHR)


0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
R^2
75
150
500
1000
2000
4000
8000
n_R

(SRR)

Fig 2. Comparison of different methods. SHL: Markers and QTL were selected for high and low MAFs, respectively; SHR: Markers were selected for high MAF, whereas QTL were selected at random; SRR: Markers and QTL were selected at random; n_R: number of individuals in RP; R^2: squared correlation between y and $\hat{u}$; simulated heritability, h^2_{sim}, (grey line); prediction R^2s of GBLUP (red line), BayesB (green line) and BayesC (purple line).

doi:10.1371/journal.pone.0161054.g002

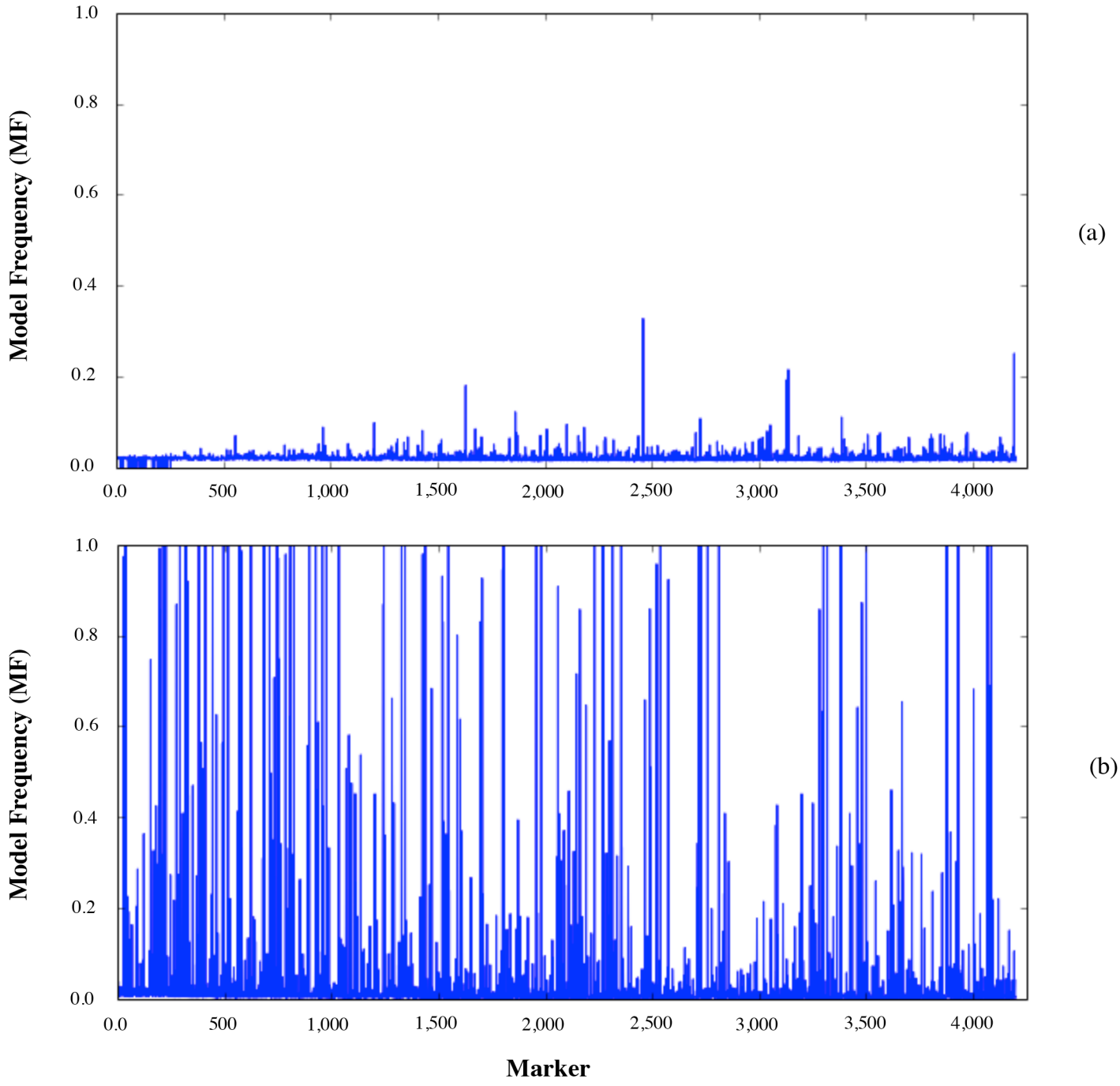


Fig 3. Model frequencies of markers in one replicate of SHL for BayesB method. SHL: Markers and QTL were selected for high and low MAFs, respectively; (a) $n_R = 75$; (b) $n_R = 8{,}000$.

doi:10.1371/journal.pone.0161054.g003

When the QTL were included in the panel, the gap between predictive accuracies of GBLUP and BayesB or BayesC was higher than when only markers were in the panel (Tables 2–4). Moreover, in this case, BayesB and BayesC with relatively small values of n_R (500-1,000) could achieve the predictive accuracy of GBLUP with n_R = 8,000.

It was reported [41] that the predictive accuracy of BayesB, was not greater than that of GBLUP, which we believe is due to the circumstance that the number of markers fitted ($\sim$2.5 million) greatly exceeded the small number of RP, n_R = 155. Several studies have investigated the effect of different methods on predictive accuracy [10–12, 24, 40, 42], and it can be concluded that none of these methods is universally best, and the performance of the method depends on the genetic architecture and the heritability of the trait as well as the RP size.

Realized Relationships and Estimability

We examined the effect of relationship on predictive ability using the maximum realized relationship of individuals in VP, $max(\mathbf{g}_{V_i})$, to the individuals in RP for two extreme groups of individuals, those with low relationship (L) or high relationship (H). Table 5 summarizes the predictive abilities obtained by GBLUP for L and H groups at varying RP sizes and for two quantitative traits with h^2 = 0.8 and 0.999. An important aspect of Table 5 is that, when RP size, n_R, increased, the number of individuals (n_L) with $max(\mathbf{g}_{V_i})<0.15$ decreased, while the number of individuals n_H with $max(\mathbf{g}_{V_i})\geq 0.25$ increased. In the scenario where h^2 = 0.8, when n_R = 500 and n_R = 5,000 there were 1,347 and 163 individuals in L group, while there were 8 and 69 individuals in H group. The trends in the number of individuals in both L and H groups indicates that when more individuals are included in RP, the probability of having at least one individual in the RP with a high relationship for a VP individual increases. However, even when n_R was 5,000, among the 2,000 VP individuals there were still only 69 and 64 of 2,000 individuals in H group for h^2 = 0.8 and h^2 = 0.999, respectively.

Predictive ability tended to increase not only for H group but also for L group individuals as the n_R increased (Table 5). For h^2 = 0.8, predictive abilities in L group, R^2_L, were 0.352, 0.477, 0.614 and 0.699, while predictive abilities in H group, R^2_H, were 0.439, 0.469, 0.614 and 0.733 when n_R was 500, 1,000, 2,000 and 5,000, respectively. For h^2 = 0.999, Predictive abilities in L group, R^2_L, were 0.542, 0.756, 0.932 and 0.995, while in H group R^2_H were 0.533, 0.704, 0.939 and 0.996 when n_R was 500, 1,000, 2,000 and 5,000, respectively. This implies that even when the

Table 5. Predictive accuracies and estimabilities for relationship groups.

h^2	n_R	n_V	n_L	n_H	R^2_L	R^2_H	$\bar{UP}_L$	$\bar{UP}_H$
0.8	500	2,000	1347	8	0.352(0.012)	0.439(0.083)	0.549(0.000)	0.569(0.003)
	1,000	2,000	972	17	0.477(0.006)	0.469(0.050)	0.777(0.000)	0.791(0.001)
	2,000	2,000	546	31	0.614(0.009)	0.614(0.043)	0.942(0.000)	0.947(0.000)
	5,000	2,000	163	69	0.699(0.011)	0.733(0.017)	1.000(0.000)	1.000(0.000)
0.999	500	2,000	1342	8	0.542(0.011)	0.533(0.080)	0.550(0.000)	0.575(0.002)
	1,000	2,000	963	15	0.756(0.005)	0.704(0.047)	0.777(0.000)	0.792(0.001)
	2,000	2,000	558	29	0.932(0.002)	0.939(0.008)	0.942(0.000)	0.948(0.000)
	5,000	2,000	166	64	0.995(0.000)	0.996(0.000)	1.000(0.000)	1.000(0.000)

h^2: heritability of the trait; n_R: number of individuals in RP; n_V: number of individuals in VP; n_L and n_H are the average number individuals in L and H groups over replicates, respectively; R^2_L and R^2_H are the predictive accuracies in L and H groups, respectively; $\bar{UP}_L$ and $\bar{UP}_H$ are the mean of the average *UP* at each replicate for L and H, respectively.

doi:10.1371/journal.pone.0161054.t005

pairwise relationships between VP and RP individuals are low, one can obtain high predictive ability.

We approached GP as an estimability problem, and derived an upper bound for reliability, and thus the upper bound for R^2. Mean values of upper bound for reliability, $\bar{UP}$, are also given in Table 5 for L and H groups. When multiplied by the trait heritability, the upper bound for R^2 is obtained, $h^2\bar{UP}$. For $h^2 = 0.8$, the upper bounds of R^2_L were 0.439, 0.622, 0.754 and 0.800, whereas the upper bounds of R^2_H were 0.455, 0.633, 0.758 and 0.800. When $h^2 = 0.999$, the upper bounds for R^2_L were 0.549, 0.776, 0.941 and 0.999, whereas the upper bounds for R^2_H were 0.574, 0.791, 0.947 and 0.999.

Our results indicate that predictive ability depends on how well an individual's genotype vector in VP can be written as a linear combination of the rows of the genotype matrix of RP individuals. As n_R increases, the row space of $\mathbf{X}_R$ will tend to increase and the possibility that $\mathbf{x}_V$ is in the row space of $\mathbf{X}_R$ will also increase. Based on these results, it can be concluded that prediction R^2 is limited by $h^2\bar{UP}$.

Habier et al. [9, 12], showed that a high relationship between the individuals in VP and RP resulted in a high predictive ability using simulated and real data. Legarra et al. [34], reported a higher predictive ability within-family than across-family in mice. Clark et al. [33], concluded that the overall prediction of breeding values relied on the degree of relationship between the VP and RPs. Pszczola et al. [13], examined the predictive ability for varying levels of relationships within RP, and between VP and RP. Their results also showed that to achieve a high predictive ability, a high relationship is required between VP and RP. Makowsky et al. [15], showed that predictive ability increases with an increase in the number of close relatives of VP individuals in the RP. On the other hand, Luan et al. [10] investigated predictive ability of GP for a dairy cattle breed, and their findings indicated an important aspect of the relationship between RP and VPs. Contrary to the above-mentioned studies, their results did not provide any strong evidence for the effect of relationship between RP and VP. In this study, we have shown that provided that the genotypes of VP individuals are in the row space of $\mathbf{X}_R$, high predictive ability can be obtained depending on the heritability of the trait and the RP size even when the pairwise relationships between VP and RP are low. This is consistent with the suggestion by Calus [43] that use of a RP comprising the whole range of phenotypes and genotypes is the requirement to obtain reliable predictions.

Supporting Information

S1 Text. Includes the derivations to reach the approximation for R^2 given as Eq (1), and derivations leading to the upper bound of R^2 presented in the manuscript.
(PDF)

Author Contributions

Conceived and designed the experiments: RLF.

Performed the experiments: EK.

Analyzed the data: EK.

Contributed reagents/materials/analysis tools: EK HC RLF DJG.

Wrote the paper: EK HC MZF DJG RLF.

Derived the upper bound for predictive ability: RLF EK HC.

References

1. Meuwissen THE, Hayes BJ, and Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157:1819–1829. PMID: 11290733
2. Goddard M and Hayes BJ (2007) Genomic selection. J Anim Breed Genet, 124:323–330. doi: 10.1111/j.1439-0388.2007.00702.x PMID: 18076469
3. Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, and Goddard ME (2009) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol, pages 41–51.
4. Wolc A, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R, Fernando R, Garrick DJ, and Dekkers JC (2013) Analysis of egg production in layer chickens using a random regression model with genomic relationships. Poult Sci, 92(6):1486–1491. doi: 10.3382/ps.2012-02882 PMID: 23687143
5. Bernardo R and Yu J (2007) Prospects for genome-wide selection for quantitative traits in maize. Crop Sci, 47:1082–1090. doi: 10.2135/cropsci2006.11.0690
6. Heffner E, Sorrells M, and Jannink J (2009) Genomic selection for crop improvement. Crop Sci, 49:1–12. doi: 10.2135/cropsci2008.08.0512
7. Jannink J-L, Lorenz AJ, and Iwata H (2010) Genomic selection in plant breeding: From theory to practice. Brief Funct Genomics, 9(166-177). PMID: 20156985
8. McCarthy JJ, McLeod HL, and Ginsburg GS (2013) Genomic medicine: A decade of successes, challenges, and opportunities. Sci Transl Med, 5:189–4. doi: 10.1126/scitranslmed.3005785
9. Habier D, Fernando RL, and Dekkers JC (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177:2389–2397. doi: 10.1534/genetics.107.081190 PMID: 18073436
10. Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, and Meuwissen THE (2009) The accuracy of genomic selection in norwegian red cattle assessed by cross-validation. Genetics, 183:1119–1126. doi: 10.1534/genetics.109.107391 PMID: 19704013
11. Daetwyler HD, Pong-Wong R, Villanueva B, and Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics, 185(3):1021–1031. doi: 10.1534/genetics.110.116855 PMID: 20407128
12. Habier D, Tetens J, Seefried FR, Lichtner P, and Thaller G (2010) The impact of genetic relationship information on genomic breeding values in german holstein cattle. Genet Sel Evol, 42(5). doi: 10.1186/1297-9686-42-5 PMID: 20170500
13. Pszczola M, Strabel T, Mulder HA, and Calus MPL (2011) Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci, 95:389–400. doi: 10.3168/jds.2011-4338
14. Clark SA, Hickey JM, Daetwyler HD, and van der Werf JH (2012) The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol, 44(4). doi: 10.1186/1297-9686-44-4 PMID: 22321529
15. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, et al. (2011) Beyond missing heritability: Prediction of complex traits. PLoS Genet, 7(4):e1002051. doi: 10.1371/journal.pgen.1002051 PMID: 21552331
16. Goddard M (2009) Genomic selection: prediction of accuracy and maximization of long term response. Genetica, 136(245-257). PMID: 18704696
17. De los Campos G, Pong-Wong R Hickey JM, Daetwyler HD, and Calus MPL (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193(2):327–345. doi: 10.1534/genetics.112.143313 PMID: 22745228
18. Fernando RL (1998) Genetic evaluation and selection using genotypic, phenotypic and pedigree information. 6th WCGALP, Armidale, Australia, 11-16 January.
19. VanRaden PM (2008) Efficient methods to compute genomic predictions J Dairy Sci, 91(11):4414–4423. doi: 10.3168/jds.2007-0980 PMID: 18946147
20. Stranden I and Garrick DJ (2009) Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci, 92(6):2971–2975. doi: 10.3168/jds.2008-1929 PMID: 19448030
21. Nejati-Javaremi A, Smith C, and Gibson JP (1997) Effect of total allelic relationship on accuracy of evaluation and response to selection. J Anim Sci, 75:1738–1745. PMID: 9222829
22. Henderson CR (1984) Applications of linear models in animal breeding. Guelph, Ontario, Canada: Univ. Guelph.
23. Hayes BJ and Goddard ME (2001) The distribution of the effects of genes affecting quantitative traits in livestock. Genet Sel Evol, 33:209–229. doi: 10.1186/1297-9686-33-3-209

24. Meuwissen THE and Goddard M (2010) Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics, 185:623–631. doi: 10.1534/genetics.110.116590 PMID: 20308278

25. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al (2010) Common SNPs explain a large proportion of the heritability for human height. Nature Genetics, 42:565–569. doi: 10.1038/ng.608 PMID: 20562875

26. Kizilkaya K, Fernando RL, and Garrick DJ (2010) Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J Anim Sci, 88:544–551. doi: 10.2527/jas.2009-2064 PMID: 19820059

27. Habier D, Fernando RL, Kizilkaya K, and Garrick DJ (2010) Extension of the bayesian alphabet for genomic selection. BMC bioinformatics, 12.

28. Daetwyler HD, Villanueva B, and Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE, 3(10):e3395. doi: 10.1371/journal.pone.0003395 PMID: 18852893

29. Goddard ME, Hayes BJ, and Meuwissen THE (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet, 128:409–421. doi: 10.1111/j.1439-0388.2011.00964.x PMID: 22059574

30. de los Campos G, Vazquez AI, Fernando RL, Klimentidis YC, and Sorensen D (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet, 9(7):e1003608. doi: 10.1371/journal.pgen.1003608 PMID: 23874214

31. Takahata N (1993) Allelic genealogy and human evolution. Mol Biol Evol, 10(1):2–22. PMID: 8450756

32. Perez-Cabal MA, Vazquez AI, Gianola D, Rosa GJM, and Weigel KA (2012) Accuracy of genome-enabled prediction in a dairy cattlepopulation using different cross-validation layouts. Frontiers in Genetics, 3(27). doi: 10.3389/fgene.2012.00027 PMID: 22403583

33. Clark SA, Hickey JM, and van der Werf HJ (2011) Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol, 43(18). doi: 10.1186/1297-9686-43-18 PMID: 21575265

34. Legarra A, Robert-Grani C, Manfredi E, and Elsen JM (2008) Performance of genomic selection in mice. Genetics, 180:611–618. doi: 10.1534/genetics.108.088575 PMID: 18757934

35. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature, 491:56–65. doi: 10.1038/nature11632 PMID: 23128226

36. Cheng H, Garrick D, and Fernando R (2015) Xsim: Simulation of descendants from ancestors with sequence data. G3: Genes|Genomes|Genetics, 5(7):1415–1417. doi: 10.1534/g3.115.016683 PMID: 25953958

37. Fernando RL and Garrick DJ (2013) Bayesian Methods Applied to GWAS. In Genome-Wide Association Studies and Genomic Prediction. Springer: Humana Press.

38. Garrick DJ and Fernando RL (2013) Implementing a QTL detection study (GWAS) using genomic prediction methodology. In Genome-Wide Association Studies and Genomic Prediction. Springer: Humana Press.

39. Scheffe H (1999) The analysis of variance, volume 72. John Wiley & Sons.

40. Meuwissen THE (2009) Accuracy of breeding values of unrelated individuals predicted by dense snp genotyping. Genet Sel Evol, 41(35).

41. Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, et al. (2011) Using whole-genome sequence data to predict quantitative trait phenotypes in drosophila melanogaster. PLoS genetics, 8(5).

42. Zhang Z, Liu J, Ding X, Bijma P, de Koning DJ, Zhang Q (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS ONE, 5(9):e12648. doi: 10.1371/journal.pone.0012648 PMID: 20844593

43. Calus MPL (2010) Genomic breeding value prediction: methods and procedures. Animal, 4(2):157–164. doi: 10.1017/S1751731109991352 PMID: 22443868